Nucleic Acids Research, 2004, Vol. 32, Database issue ***D505–D508***
DOI: 10.1093/nar/gkh022

LIFEdb: a database for functional genomics experiments integrating information from external sources, and serving as a sample tracking system

Detlev Bannasch*, Alexander Mehrle, Karl-Heinz Glatting[1], Rainer Pepperkok[2], Annemarie Poustka and Stefan Wiemann

German Cancer Research Center (DKFZ), Division of Molecular Genome Analysis, Im Neuenheimer Feld 580, D-69120 Heidelberg, Germany, [1]German Cancer Research Center (DKFZ), Division of Molecular Biophysics, Im Neuenheimer Feld 580, D-69120 Heidelberg, Germany and [2]European Molecular Biology Laboratory (EMBL), Department of Cell Biology and Cell Biophysics, Meyerhofstrasse 1, D-69117 Heidelberg, Germany

Recieved August 1, 2003; Revised and Accepted September 3, 2003

ABSTRACT

We have implemented LIFEdb (http://www.dkfz.de/LIFEdb) to link information regarding novel human full-length cDNAs generated and sequenced by the German cDNA Consortium with functional information on the encoded proteins produced in functional genomics and proteomics approaches. The database also serves as a sample-tracking system to manage the process from cDNA to experimental read-out and data interpretation. A web interface enables the scientific community to explore and visualize features of the annotated cDNAs and ORFs combined with experimental results, and thus helps to unravel new features of proteins with as yet unknown functions.

INTRODUCTION

Although the whole sequence of the human genome has been determined and ~30 000 genes have been predicted (1,2), the function of many of these genes and their encoded proteins remains unknown. To address this topic efficiently, various high-throughput methods are being developed. As a first step, full-length cDNA sequencing projects have been established in order to identify cDNA clones that contain a complete open reading frame (ORF) (3,4). These clones serve as starting material to set up functional experiments dependent on full-length proteins. While use of cDNAs on a small scale has been carried out for a long time, there are currently only a few functional genomics and proteomics experiments that systematically exploit cDNA resources on a larger scale (5,6).

Determining the subcellular localization of novel proteins has proved to be a valuable tool for gaining insight into the possible function of a protein, since a protein's localization determines its microenvironment and potential interaction partners (5,7,8). Data on the subcellular localization patterns of yeast proteins (9,10) and of murine nuclear proteins (11) are available in online databases.

Here, we describe the Database for Localization, Interaction, Functional assays and Expression of Proteins (LIFEdb), which integrates data regarding novel full-length cDNAs generated by the German cDNA Consortium and data produced during the cloning process of selected ORFs, as well as data on the expression of the encoded proteins and their subcellular localization in a mammalian cell line. The database also serves as a sample-tracking system to easily monitor the whole plasmid construction process. Furthermore, a web interface is provided which enables the scientific community to explore significant features of annotated cDNAs and ORFs and to combine them with experimental results. The combination of these various data will help the discovery of new functions of the encoded proteins.

BACKGROUND

To carry out expression studies of novel full-length proteins, selected ORFs are systematically amplified by PCR and cloned by recombination into an entry vector from which they can be shuttled into Gateway (Invitrogen)-compatible expression vectors for eukaryotic and prokaryotic protein expression. To systematically determine the subcellular localization of the novel proteins, the ORFs are subcloned into EYFP and ECFP expression vectors for the generation of C-terminal yellow fluorescent and N-terminal cyan fluorescent fusion proteins, respectively (5). Performing this task on a larger scale requires, in addition to the cloning system being amenable to high throughput, software that (i) helps to keep track of the wet-lab processes in the cloning and expression processes, (ii) stores the annotation and experimental results and (iii) makes the data available to researchers for integrated analysis.

*To whom correspondence should be addressed. Tel: +49 6221 424711; Fax: +49 6221 423454; Email: d.bannasch@dkfz.de
Correspondence may be also addressed to Alexander Mehrle. Tel: +49 6221 424711; Fax: +49 6221 423454; Email: a.mehrle@dkfz.de

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

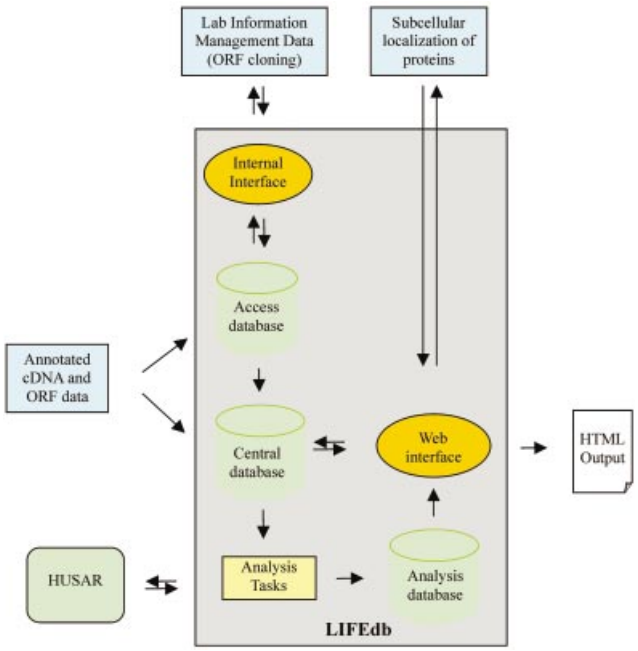

Figure 1. Overview of the data flow in LIFEdb. Annotated data from novel full-length cDNAs are entered into both a MS Access database and into the central database, which is based on MS SQL Server 2000. The Access database serves as a laboratory information management system, which is accessible via an internal interface based on Access-derived forms. The data within this database is regularly copied into the central database. LIFEdb also provides a web interface, which is accessible to both the scientific community and via a password-protected area to remote users who retrieve data regarding expression clones, and who in turn can upload experimental results, including pictures of the subcellular localizations. For each cDNA in the central database, a comprehensive bioinformatic analysis of the deduced protein sequences is performed using programs and tasks of the HUSAR system. The results are then stored in the analysis database. The web interface allows for retrieval of these data and presents them, together with annotation data and data from the subcellular localization experiments, to the scientific community.

DATABASE DESIGN

LIFEdb (Fig. 1) contains an MS Access database for storage of both selected annotation data of the cDNA clones used and data produced in the subsequent cloning processes. The latter data are entered via an internal interface consisting of MS Access-derived forms. Furthermore, LIFEdb contains a central database and an analysis database, both of which are based on MS SQL Server 2000. The central database is a regularly updated mirror of the MS Access database, and in addition it stores further annotation data and the results of the subcellular localization experiments. The analysis database is used for storage of data produced from bioinformatic analysis of the proteins encoded by the cDNAs under investigation.

Via a web interface based on MS Internet Information Server (IIS), selected cDNA and protein data (consisting of the annotation, experimental and analysis data) can be retrieved by the scientific community (Fig. 2). Furthermore, the web interface contains a password-protected area that allows data exchange with collaborators outside the DKFZ.

In the initial step, data from annotated cDNA sequences and corresponding ORF information are stored in LIFEdb. cDNA analysis is performed to some extent automatically using tools available from the internet and from the HUSAR system (http://genome.dkfz-heidelberg.de/) (12,13). Subsequently, annotation is carried out manually (I. Schupp and V. Kuryshev, personal communication).

Data regarding the cloning procedure, which comprises primer and PCR data, as well as data concerning the construction process of entry and destination vectors are entered into the Access database via the internal interface. Various Visual Basic for Application (VBA) programs are running in the background of the internal interface. They enable the user to keep track of the whole cloning process easily and, together with the database system itself, they ensure the consistency of the data.

Determination of the subcellular localization of the proteins is performed outside the DKFZ (R. Pepperkok's group at the EMBL). Therefore, it was necessary to implement a web interface for the uploading of raw data and information, which allows for a decentralized production of data, while still keeping the data centralized. The web interface enables the remote researcher to retrieve selected experimental data produced in our group (including data concerning the ECFP and EYFP destination clones) from the central database in LIFEdb and to enter their results from subcellular localization experiments, together with comments and experimental conditions. Furthermore, primary data such as microscopic images showing the subcellular localization of the investigated proteins can be uploaded.

Currently, LIFEdb contains the data of more than 1100 full-length ORFs and the results of subcellular localization experiments of more than 500 different proteins.

AUTOMATED PROTEIN ANALYSIS

For each cDNA in the database, a comprehensive bioinformatic analysis of the deduced protein sequences is performed automatically (C. del Val, manuscript in preparation) utilizing programs and tasks of the HUSAR (Heidelberg UNIX Sequence Analysis Resources) system (http://genome.dkfz-heidelberg.de/) (12,13). The analysis includes the determination of basic physicochemical parameters (e.g. molecular weight, isoelectric point) of the encoded protein, the prediction of its subcellular localization [PSort (14)], a search for motifs [e.g. signal sequences, prosite patterns/profiles (15)] and a search against the Swissprot and Trembl protein databases (16) using the BlastP2 program (17). The data from these bioinformatic analyses are presented via the web interface (Fig. 2D) and are updated regularly (e.g. BlastP2 alignments are calculated on a weekly basis).

SEARCHING THE DATABASE

The web interface allows for database searches to obtain information about the cDNAs, encoded proteins and associated experimental results. The home page contains an input field, which carries out a text-based analysis and recognizes accession numbers, clone IDs and various keywords, which are then searched for. The results are presented as a table showing selected features of the cDNA(s) and the

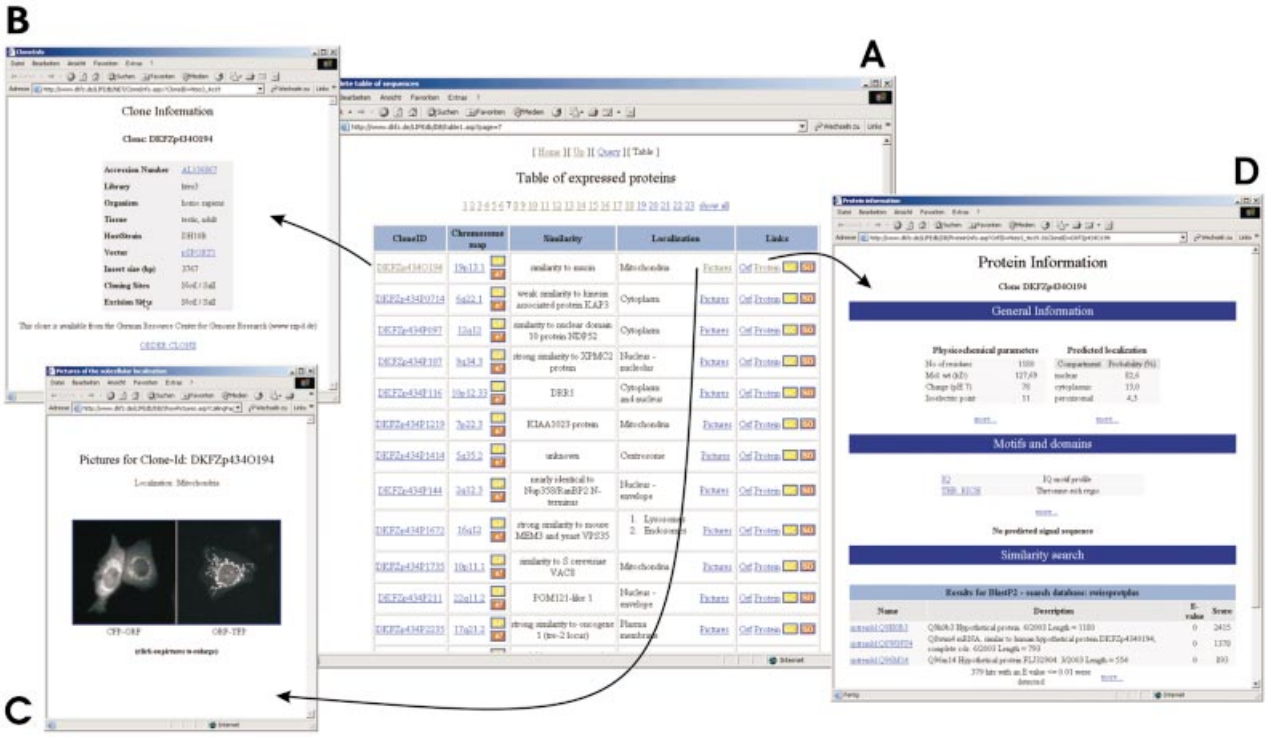

Figure 2. Example of a database search using the web interface of LIFEdb. Selected information about the cDNAs, the encoded proteins and experimental results stored in LIFEdb are presented as a table, which in addition contains links to external databases (Ensembl, GoldenPath and GeneCards) (**A**). Table entries are linked to pages with additional data such as general information on a selected cDNA clone (**B**), representative pictures of the subcellular localization of the encoded protein (**C**) and a summary of the bioinformatic analysis of the protein (**D**).

corresponding ORF(s) together with the subcellular compartment(s) in which the expressed protein(s) are localized (Fig. 2A). Table entries are linked to pages with additional data, such as general information on a selected cDNA clone (Fig. 2B), representative pictures of the subcellular localization of the encoded protein (Fig. 2C) and a summary of the bioinformatic analysis of the protein (Fig. 2D). Furthermore, links to external databases such as Ensembl (18), GoldenPath (19), GeneCards (20) (Fig. 2A) and to the sequence entry of the EMBL database (21) are provided (Fig. 2B). An additional query page allows for more criteria to be set (e.g. chromosomal locus, subcellular compartment). Taken altogether, LIFEdb combines experimental data and data of external data sources by presenting them in a web interface and hence supports the discovery of new functions of as yet uncharacterized or only partially characterized proteins under investigation.

PROSPECT

Future developments will involve the implementation of the results of functional experiments currently under development into LIFEdb and their retrieval via the web interface. Furthermore, the automatic protein analysis will be extended to incorporate a more comprehensive motif and domain search as well as a prediction of the secondary structure of the proteins investigated.

ACKNOWLEDGEMENTS

We thank Patricia McCabe and Lee Bergman for critical reading of the manuscript and Stefanie Bechtel, Anni Duda, Kerstin Hettler and Heike Wilhelm for suggestions on the internal interface. This work is supported by the Bundesministerium für Bildung und Forschung within the DHGP (01KW0012) and NGFN (01GR0101) projects.

REFERENCES

1. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. International Human Genome Sequencing Consortium. *Nature*, **409**, 860–921.
2. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
3. Nagase,T., Nakayama,M., Nakajima,D., Kikuno,R. and Ohara,O. (2001) Prediction of the coding sequences of unidentified human genes. XX. The complete sequences of 100 new cDNA clones from brain which code for large proteins *in vitro*. *DNA Res.*, **8**, 85–95.
4. Wiemann,S., Weil,B., Wellenreuther,R., Gassenhuber,J., Glassl,S., Ansorge,W., Bocher,M., Blocker,H., Bauersachs,S., Blum,H. *et al.* (2001) Toward a catalog of human genes and proteins: sequencing and analysis of 500 novel complete protein coding human cDNAs. *Genome Res.*, **11**, 422–435.
5. Simpson,J.C., Wellenreuther,R., Poustka,A., Pepperkok,R. and Wiemann,S. (2000) Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO Rep.*, **1**, 287–292.
6. Ziauddin,J. and Sabatini,D.M. (2001) Microarrays of cells expressing defined cDNAs. *Nature*, **411**, 107–110.

7. Pepperkok,R., Simpson,J. and Wiemann,S. (2001) Being in the right location at the right time. *Genome Biol.*, **2**, REVIEWS1024.
8. Sutherland,H.G., Mumford,G.K., Newton,K., Ford,L.V., Farrall,R., Dellaire,G., Caceres,J.F. and Bickmore,W.A. (2001) Large-scale identification of mammalian proteins localized to nuclear sub-compartments. *Hum. Mol. Genet.*, **10**, 1995–2011.
9. Habeler,G., Natter,K., Thallinger,G.G., Crawford,M.E., Kohlwein,S.D. and Trajanoski,Z. (2002) YPL.db: the Yeast Protein Localization database. *Nucleic Acids Res.*, **30**, 80–83.
10. Kumar,A., Cheung,K.H., Ross-Macdonald,P., Coelho,P.S., Miller,P. and Snyder,M. (2000) TRIPLES: a database of gene function in *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **28**, 81–84.
11. Dellaire,G., Farrall,R. and Bickmore,W.A. (2003) The Nuclear Protein Database (NPD): sub-nuclear localisation and functional annotation of the nuclear proteome. *Nucleic Acids Res.*, **31**, 328–330.
12. Ernst,P., Glatting,K.H. and Suhai,S. (2003) A task framework for the web interface W2H. *Bioinformatics*, **19**, 278–282.
13. Senger,M., Glatting,K.H., Ritter,O. and Suhai,S. (1995) X-HUSAR, an X-based graphical interface for the analysis of genomic sequences. *Comput. Methods Programs Biomed.*, **46**, 131–141.
14. Nakai,K. and Horton,P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, **24**, 34–36.
15. Sigrist,C.J., Cerutti,L., Hulo,N., Gattiker,A., Falquet,L., Pagni,M., Bairoch,A. and Bucher,P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.*, **3**, 265–274.
16. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledge base and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
17. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
18. Clamp,M. andrews,D., Barker,D., Bevan,P., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V. *et al.* (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.*, **31**, 38–42.
19. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.
20. Safran,M., Chalifa-Caspi,V., Shmueli,O., Olender,T., Lapidot,M., Rosen,N., Shmoish,M., Peter,Y., Glusman,G., Feldmesser,E. *et al.* (2003) Human gene-centric databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. *Nucleic Acids Res.*, **31**, 142–146.
21. Stoesser,G., Baker,W., van den Broek,A., Camon,E., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V. *et al.* (2002) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **30**, 21–26.